# Chapter 5

# Similarity, Equivalence Testing, and Discrimination Theory

**Abstract** This chapter discusses equivalence testing and how difference tests are modified in their analyses to guard against Type II error (missing a true difference). Concepts of test power and required sample sizes are discussed and illustrated. An alternative approach to equivalence, namely interval testing is introduced along with the concept of paired one-sided tests. Two theoretical approaches to the measurement of the size of a difference are introduced: discriminator theory (also called guessing models) and the signal detection or Thurstonian models.

*Difference testing method constitute a major foundation for sensory evaluation and consumer testing. These methods attempt to answer fundamental questions about stimulus and product similarity before descriptive or hedonic evaluations are even relevant. In many applications involving product or process changes, difference testing is the most appropriate mechanism for answering questions concerning product substitutability.*

—D. M. Ennis (1993)

## Contents



## 5.1 Introduction

Discrimination, or the ability to differentiate two stimuli, is one of the fundamental processes underlying other sensory-based responses. As suggested by Ennis above, if two items cannot be discriminated, there is no basis for description of differences between them, nor for consumer preference. The previous chapter discussed simple discrimination tests that are used to gather evidence that a product has changed from some previous version. We might make an ingredient change, a cost reduction, a processing or packaging change, do a shelf life test against a fresh control, or a quality control test against some standard product. Questions arise as to whether the difference in

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_5, 

the products is perceivable. Discrimination or simple difference tests are appropriate for these questions. When we find evidence of a difference, the methods are straightforward and the interpretation is usually clear-cut as well. However, a great deal of sensory testing is done in situations where the critical finding is one of equivalence or similarity. That is, a no-difference result has important implications for producing, shipping, and selling our product. Shelf life and quality control tests are two examples. Cost reductions and ingredient substitutions are others.

This is a much trickier situation. It is often said that "science cannot prove a negative" and the statistical version of this is that you cannot really prove that the null hypothesis is correct. But in a way, this is exactly what we are trying to do when we amass evidence that two products are equivalent or sufficiently similar that we can substitute one for the other without any negative consequences.

The issue is not so easy as just finding "no significant difference." A failure to reject the null is always ambiguous. Just because two products are "not significantly different" does not necessarily mean that they are equivalent, sensorially. There are a many reasons why we may have failed to find a statistically significant difference. We may not have tested enough people relative to the amount of error variability in our measurements. The error variability may be high for any number of reasons, such as lack of sample control, poor testing environment, unqualified judges, poor instructions, and/or a bad test methodology. It is easy to do a sloppy experiment and get a non-significant result.

Students and sensory scientists should recall that there are two kinds of statistical errors that can be made in any test. These are shown in Fig. 5.1. We can reject the null and conclude that products are different when they are not. This is our familiar Type I error that we try to keep to a long-term minimum called our alpha level. This is commonly set at 5%, and why we try to use probability levels of 0.05 as our cutoffs in statistical significance testing. This kind of error is dangerous in normal experimental science and so it is the first kind of error that we worry about. If a graduate student is studying a particular effect, but that effect was a spurious false-positive result from some earlier test, then he or she is wasting time and resources. If a product developer is trying to make an improved product, but that hunch is based on an earlier false result, the effort is doomed. Some people refer to Type I error as a "false alarm." The second kind of error occurs when we miss a difference that is really there. This is a Type II error, when we do not reject the null, but the alternative hypothesis is really the true state of the situation.

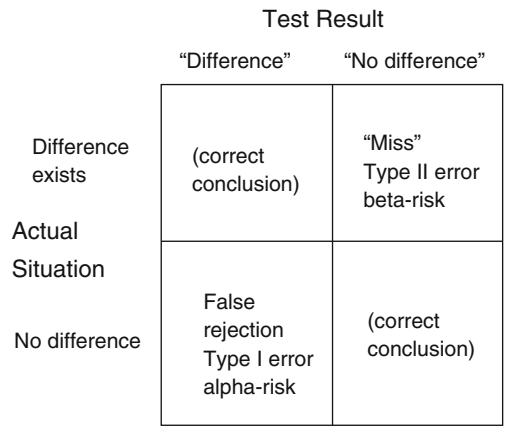



**Fig. 5.1** The statistical decision matrix showing the two types of error, Type I, when the null is rejected but there is really no difference, and Type II, when there is a difference but none is detected by the test (null false but accepted). The long-term risk of Type I under a true null is the alpha risk. Beta risk is managed by the choices made in the experiment of *N*, alpha, and the size of the difference one is trying to detect.

Type II error has important business ramifications, including lost opportunities and franchise risk. We can miss an opportunity to improve our product if we do not find a significant effect of some ingredient or processing change. We can risk losing market share or "franchise risk" if we make a poorer product and put it into the marketplace because we did not detect a negative change. So this kind of error is of critical importance to sensory workers and managers in the foods and consumer products industries.

The first sections of this chapter will deal with ways to gather evidence that two products are similar or equivalent from a sensory perspective. The first section will illustrate some commonsense approaches and the question of test power. Then some formal tests for similarity or equivalence will be considered, after we look at a model for estimating the size of a difference based on the proportion of people discriminating.

After discussing basic approaches to similarity and equivalence, this chapter will examine more sophisticated models for measuring sensory differences from discrimination results. Sensory professionals need to do more than simply "turn the crank" on routine tests and produce binary yes/no decisions about statistical

significance. They are also required to understand the relative sensitivity of different test methods, the decision processes and foibles of sensory panelists, and the potential pitfalls of superficial decisions. For these reasons, we have included in this chapter sections on the theory of signal detection and its related model, Thurstonian scaling. Many questions can arise from apparently simple discrimination tests in applied research, as the following examples show:

(1) Clients may ask, "OK, you found a difference, but was it a big difference?"
(2) When is it acceptable to conclude that two products are sensorially equivalent when the test says simply that I did not get enough correct answers to "reject the null?"
(3) What can I do to insure that the test is as sensitive as possible and does not miss an important difference (i.e., protect against Type II error)?
(4) Why are some test methods more stringent or more difficult than others? Can this be measured?
(5) What are the behaviors and decision processes that influence sensory-based responses?

Each of these questions raise difficult issues without simple answers. This chapter is designed to provide the sensory professional with approaches to these questions and a better understanding of methods, panelists, and some enhanced interpretations of discrimination results. For further detail, the books by Bi (2006a) on discrimination testing, by Welleck (2003) on equivalence testing, and the general sensory statistics book by Gacula et al. (2009) are recommended.

## 5.2 Common Sense Approaches to Equivalence

Historically, many decisions about product equivalence were based on a finding of no significant difference in a simple discrimination test. This is a risky enterprise. If the difference is subtle, it is easy to miss it. The test might have included too few panelists for the size of the effect. We may have let unexpected sources of variability creep into the test situation, ones that could easily overwhelm the difference. We may have used unqualified panelists because too many of our regular testers were on vacation or out sick that week. Perhaps our sample-handling equipment like heat lamps were not working that day. Any number of reasons could contribute to a sloppy test that would create situation in which a difference could be missed. So why was this approach so prevalent during the early history of sensory testing?

There are some common sense situations in which it may make sense to consider a non-significant result as important. The first requirement is that the test instrument must be proven to detect differences in previous tests. By "test instrument" we mean the entire scenario – a particular method, a known population of panelists, specific test conditions, these type of products, etc. In a company with an ongoing test program this kind of repeated scenario may provide a reasonable insurance policy that when a significant difference is not found, it was in fact not due to a faulty instrument, as the instrument has a track record. For example, a major coffee company may have ongoing tests to insure that the blend and roasting conditions produce a reliable, reproducible flavor that the loyal customers will recognize as typical and accept. Other controls may be introduced to further demonstrate the effectiveness of the test method. Known defective or different samples may be given in calibration tests to demonstrate that the method can pick up differences when they are in fact expected. Conversely, known equivalents or near duplicates may be given to illustrate that the method will not result in an unacceptable rate of false alarms. Finally, the panel may consist of known discriminators who have been screened to be able to find differences and who have a track record of detecting differences that are confirmed by other tests on the same samples, such as consumer tests or descriptive analysis panels.

These kinds of controls are illustrated in a paper on cross-adaptation of sweet taste materials (Lawless and Stevens, 1983). In cross-adaptation studies, exposure to one taste substance may have an adapting effect, i.e., cause a decrease in intensity of a second substance. To claim that there is full cross-adaptation, the decrease must be about the same as with self-exposure, i.e., the decrement should be the same as when the substance follows itself. To claim no cross-adaptation, the test substance must have an intensity equivalent to its taste after plain water adaptation. Both of these are essentially equivalence tests. In order to accept these results it is necessary to prove that the second or test item is capable of being adapted (for example, it can adapt itself) and that the first-presented substance in fact can

cause an adaptation-related decrease (i.e., it also has an effect on itself). Given these two control situations, the claim of cross-adaptation (or lack thereof) as an equivalence statement becomes trustworthy.

Simple logical controls can be persuasive in an industrial situation in which there is an ongoing testing program. Such an ongoing testing program may be in place for quality control, shelf-life testing, or situations in which a supplier change or ingredient substitution is common due to variable sources of raw materials. This kind of logic is most applicable to situations in which the conditions of testing do not vary. If there is a sudden decrease in the panel size during a vacation period, for example, it becomes more difficult to claim that "we have a good instrument" and therefore a non-significant difference is trustworthy. All the testing conditions, including the panel size, must remain fairly constant to make such a claim.

## 5.3 Estimation of Sample Size and Test Power

A more statistical approach to equivalence is to manage the sample size and the test power to minimize the probability of a Type II error, i.e., the chance of missing a true difference. There are commonly accepted formulas for calculating the required sample size for a certain test power. At first glance, this seems rather straightforward. However, there is one important part of the logic that managers (and often students) find troubling. One must specify not only the acceptable amount of alpha- and beta-risks in these calculations, but also *the size of the difference one is trying to detect.* Conversely, how much of a difference would one allow, and still call the two products "equivalent" on a sensory basis? Managers, when faced with this question, may reply that they want no difference at all, but this is unrealistic and not possible within the constraints of statistical testing. Some degree of difference, no matter how minor, must be specified as a lower tolerable limit.

The common equation for calculating the necessary sample size is given as follows (from Amerine et al., 1965):

$$N = \left[\frac{Z_\alpha\sqrt{p_o q_o} + Z_\beta\sqrt{p_a q_a}}{p_o - p_a}\right]^2 \qquad (5.1)$$

where $Z_\alpha$ and $Z_\beta$ are the $Z$-scores associated with the chosen levels of alpha- and beta-risk, $p_o$ is the chance probability in the test and $p_a$ is the probability chosen for the alternative hypothesis (as always, $q = 1–p$). This is the equation for determining the required panel size, $N$, as a function of the alpha-risk, beta-risk, the chance probability level, and the effect size one does not want to miss. A similar equation (see Appendix at the end of this chapter) is used for scaled data, in which the degree of difference can be specified as a difference on a scale or number of standard deviations.

The effect size or size of the allowable difference is given in the denominator. It is this quantity that management must choose in order to determine what is sufficiently "equivalent." Strategically, management may not want to go out on a limb and may delegate this choice to the statisticians or sensory personnel involved in the program, so the sensory professional must be prepared to make recommendations based on prior experience with the product. Knowledge of the degree of consumer tolerance for differences is key in making any such recommendation. In a vacuum of consumer information, this choice can be exceedingly difficult.

In this case the size of the difference is given by stating some percentage of correct choices that is higher than the chance level, noted as $p_a$ in Eq. (5.1). You can think of this as a percent correct associated with an alternative hypothesis, or simply as a percent correct that one would not want to miss detecting. Above this level, there are too many people detecting the difference for our management level of comfort with the product change. In the next section we will introduce a useful way to think about this level, in terms of the proportion of people detecting the difference. This proportion is different than the actual observed percent correct, because we have to apply a correction for chance, i.e., the possibility that some people get the correct answer just by guessing correctly. More on this below.

For now, let us examine two worked examples for a triangle test. In the first example, we will set alpha and beta at 5%, so our one-tailed $z$-values will both be 1.645. Let us allow a fairly liberal alternative hypothesis percentage of 2/3 correct or 66.7%. This might be considered a fairly large difference, as the proportion of people truly detecting, after the correction for chance, is 50%. In other words, we might expect half the population to detect this change. On the other hand,

50% detection is considered one definition of a threshold (see Chapter 6), so from that perspective this might be an acceptable difference.

Working through the math, we get the following equation and result:

$$\left[\frac{1.645\sqrt{(0.33)(0.67)}+1.645\sqrt{(0.67)(0.33)}}{0.33-0.67}\right]^2 = 21.6$$

So for this kind of test, looking for what some managers might call a "gross difference" we need about 22 panelists. Now let us see what happens when we make the difference smaller. In this example we can only allow a correct performance level of 50% (which corresponds to 25% true detection of the difference after correction for chance or one person in four actually seeing the difference). The new equation is

$$\left[\frac{1.645\sqrt{(0.33)(0.67)}+1.645\sqrt{(0.50)(0.50)}}{0.33-0.50}\right]^2 = 91.9$$

So now that we have lowered the size of the allowable difference a little, the required panel size has expanded to 92 judges. This would be a fairly large triangle test panel by most common industrial standards. Unfortunately if you are trying to get evidence for sensory equivalence and you can permit only a small difference, you are going to need a lot of panelists! There is just no way around it, unless one goes to replicated measures and a beta-binomial approach as discussed in the previous chapter. The power of difference tests with small panels can be alarmingly low, as discussed in Appendix E. Further calculations and tables for triangle and duo–trio tests are found there as well. The important factor to note in our two examples is that it is the size of the difference as specified in the denominator of Eq. (5.1) that has the biggest influence on the panel size requirements. In the next section, we will examine a simple traditional way of choosing our acceptable size of a difference based on the estimated proportion of panelists discriminating.

## 5.4 How Big of a Difference Is Important? Discriminator Theory

How can we calculate some true measure of discrimination after adjustment for the chance level of performance? That is, a certain percent correct is expected by guessing alone in the face of no discrimination at all. An old historical approach to this was to provide a correction for the guessing level, i.e., the level of performance expected in the face of no discrimination whatsoever. The formula for the corrected proportion is known as Abbot's formula (Finney, 1971), and is given by

$$UpperC.I.95\% = \left[1.5(X/N)-0.5\right]+1.645(1.5)\sqrt{\frac{(X/N)(1-X/N)}{N}} \quad (5.2)$$

where $P_{\text{observed}}$ is the observed proportion correct and $P_{\text{chance}}$ is the proportion expected by chance guessing.

This formula has been widely applied since the 1920's in fields such as pharmacology, toxicology, and even educational testing. In pharmacology, it is used to adjust for the size of the placebo effect, i.e., those test subjects that improve without the drug. In toxicology it is used to adjust for the baseline fatality rate in the control group (i.e., those not exposed to the toxin but who die anyway). This formula has also been employed in educational testing where multiple-choice tests are common, but adjustment for guessing is desired. Another version of this formula appears in publications discussing the issue of estimating true discriminators in the sample and separating them from an estimated proportion of people who are merely guessing correctly (e.g., Morrison, 1978). The formula will also be used in the next chapter when forced-choice methods are used in threshold determinations. Chance-adjusted discrimination was unfortunately discussed initially as "recognition" in the early sensory literature (Ferdinandus et al., 1970) but we will stick with the terms discrimination and discriminators here. "Recognition" in the psychological literature implies a match to something stored in memory and that is not really the issue in discriminating a difference among samples.

The model is simple but it embraces two assumptions. The first assumption states that there are two kinds of panelists during a particular test—discriminators, who see the true difference and select the correct item and non-discriminators who see no difference and guess. The second assumption contains the logical notion that non-discriminators include people who guess correctly and those who guess incorrectly. The best estimate of the proportion guessing correctly is the chance performance level. Thus the total number

of correct judgments comes from two sources: People who see the difference and answer correctly and those who guess correctly.

In forced choice threshold measures (see Chapter 6) 50% correct performance *after adjustment for chance* is taken as a working definition (Antinone et al., 1994; Lawless, 2010; Morrison, 1978; Viswanathan et al., 1983). For example, in the triangle test or a three-alternative forced choice test, the chance level is 1/3, so 50% above chance or 50% adjusted for chance is 66.7% correct. If a paired test or duo–trio was employed, the 50% chance level now requires a 75% correct discrimination to be at threshold, when threshold is defined as a 50% correct proportion after adjustment. Another approach is to work backward, i.e., try to find the percent correct that one would expect given a targeted proportion of discriminators in the test. This is given by the re-arrangement of Abbott's formula as follows:

$$P_{\text{observed}} = P_{\text{adjusted}} + P_{\text{chance}}(1 - P_{\text{adjusted}}) \quad (5.3)$$

So for our threshold example, if we had a 3-AFC test and we required 50% discriminators, we would expect one-third of the remaining (i.e., $1–P_{\text{adjusted}}$) non-discriminators to guess correctly, thus adding 1/6 (or 1/3 of 0.5) to the 50% who see the difference and giving us 66.7% correct.

This discrimination ability should be viewed as momentary. It is not necessarily a stable trait for a given judge. That is, a given judge does not have to "always" be a discriminator or a non-discriminator. Furthermore, we are not attempting to determine who is a discriminator, we are only estimating the likely proportion of such people given our results. This is an important point that is sometimes misinterpreted in the sensory literature. A sensory panel leader who is accustomed to screening panelists to determine if they have good discriminating ability may view "discriminators" as people with a more or less stable ability and classify them as such. This is not the point here. In the guessing models, the term "discriminator" is not used to single out any individuals, in fact there is no way of knowing who was a discriminator, nor is there any need to know in order to apply the model. The model simply estimates the most likely proportion of people who are momentarily in a discriminating state and thus answer correctly as opposed to those who might be answering correctly by chance. In other words, the issue is how often people were likely to have noticed the difference.

If we choose to think about numbers correct rather than proportions, we can use a simple translation of Abbott's formula. How are the numbers of discriminators and non-discriminators estimated? The best estimate of the number of non-discriminators who guess correctly is based on the chance performance level. Once again, the total number of correct choices by the panel reflects the sum of the discriminators plus the fraction of the non-discriminators who guess correctly. This leads to the following equations: Let $N =$ number of panelists, $C =$ the number of correct answers, and $D =$ the number of discriminators. For a triangle test, the following relationships should hold:

$$C = D + \frac{1}{3}(N - D) \quad (5.4)$$

This is simply a transformation of Abbott's formula (as in Eq. (5.3)), expressed in numbers observed rather than proportions.

Here is an example: Suppose we do a triangle test with 45 judges and 21 choose the odd sample correctly. We conclude that there is a significant difference at $p < 0.05$. But how many people were actually discriminators? In this example, $N = 45$ and $C = 21$. Solving for $D$ in Eq. (5.4):

$$21 = D + \frac{1}{3}(45 - D) = \frac{2}{3}D + 15$$

and thus $D = 9$.

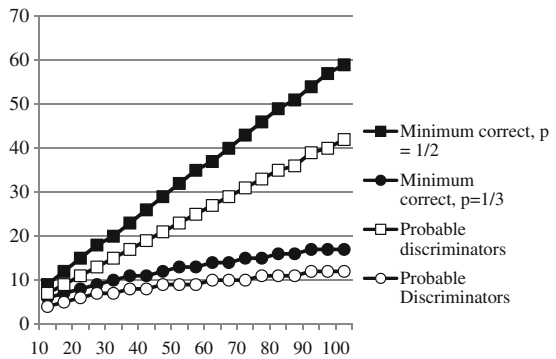
In other words, our best estimate is that 9 people out of 45 (21% of the sample) were the most likely number to have actually seen the difference. Note that this is very different from the percentage correct or 21/45 (=47%). Framed this way, a client may view the sensory result from quite a different perspective and one that is potentially more useful than the raw percent correct.

Table 5.1 shows how the number of required discriminators for various tests increases as a function of sample size. Of course, as the number of judges increases, it takes a smaller and smaller proportion of correct responses to exceed our minimum level above chance for statistical significance. This is due to the fact that our confidence intervals around the observed proportions shrink as sample size increases. We are simply more likely to have estimated a point nearer to the true population proportion correct. The number of judges getting a correct answer will also need to increase, as shown in the table. However, the number of discriminators increases at a slower rate.

**Table 5.1** Number correct versus estimated discriminators

| $N$ | Minimum correct, $p = 1/2$ | Estimated number discriminating | Minimum correct, $p = 1/3$ | Estimated number discriminating |
|---|---|---|---|---|
| 10 | 9 | 6 | 7 | 4 |
| 15 | 12 | 7 | 9 | 5 |
| 20 | 15 | 8 | 11 | 6 |
| 25 | 18 | 9 | 13 | 7 |
| 30 | 20 | 10 | 15 | 7 |
| 35 | 23 | 11 | 17 | 8 |
| 40 | 26 | 11 | 19 | 8 |
| 45 | 29 | 12 | 21 | 9 |
| 50 | 32 | 13 | 23 | 9 |
| 55 | 35 | 13 | 25 | 9 |
| 60 | 37 | 14 | 27 | 10 |
| 65 | 40 | 14 | 29 | 10 |
| 70 | 43 | 15 | 31 | 10 |
| 75 | 46 | 15 | 33 | 11 |
| 80 | 49 | 16 | 35 | 11 |
| 85 | 51 | 16 | 36 | 11 |
| 90 | 54 | 17 | 39 | 12 |
| 95 | 57 | 17 | 40 | 12 |
| 100 | 59 | 17 | 42 | 12 |

Minimum correct gives the required number for significance at $p < 0.05$, one tailed




For large sample sizes, we only need a small proportion of discriminators to push us over the statistically critical proportion for significance. Another important message for clients and management here is that although we may have found a significant difference, *not everyone can discern it.*

This way of looking at difference tests has several benefits but one serious shortcoming. One advantage is that this concept of "the proportion of discriminators" gives management an additional point of reference for interpretation of the meaning of difference tests. Statistical significance is a poor reference point, in that the significance of a given proportion correct also depends upon the number of judges. As $N$, the number of judges increases, the minimum proportion correct required for statistical significance gets smaller and smaller, approaching a level nearer to chance. So statistical significance, while providing necessary evidence against a chance result, is a poor yardstick for business decisions and is only a binary choice. The estimated proportion of discriminators is not dependent upon sample size (although confidence intervals around it are).

Another advantage to this model is that it provides a yardstick for setting panel size and a point of reference for the test of significant similarity, as outlined below. In determining a desired panel size, the experimenter must make a decision about the size of the alternative hypothesis that must be detected if it is true. That is, how much of a difference do we want to be sure to detect if it in fact exists? The correction for guessing provides one such benchmark for these calculations. Once we have decided upon a critical proportion of discriminators, we can calculate what proportion correct would be expected from the addition of (chance) guessers. This proportion correct becomes the alternative hypothesis proportion in Eq. (5.1). In other words, the choice of the alternative hypothesis can now be based on the observed proportion required to give a certain level of discriminators. We merely have to apply Abbott's formula to see what percent correct is required.

The choice should consider a strategy based on the level of normal product variability and what consumers will expect. In one situation, there might be strong brand loyalty and consumers who demand high product consistency. In that case a low proportion of discriminators might be desired in order to make a process change or an ingredient substitution. Another product might have commonplace variation that is tolerated by consumers, like some fruit or vegetable commodities, or wines from different vintages. In this case a higher proportion of discriminators might be tolerated in a difference test. As we will discuss next, in the statistical test for significant similarity, the proportion of discriminators must be decided upon in order to test for performance below that critical level, as evidence for a difference too small to be practically significant.

Let us consider a triangle test with 90 judges and 42 choose the odd sample correctly. According to Table L, this is a significant difference at $p < 0.01$. Working with Eq. (5.2), we find that the proportion of discriminators is about 20% (42/90–1/3)/(1–1/3) = 0.20. So one-fifth of the test group is our best estimate here of the proportion discriminating. For a product with an

extremely brand-loyal user group, this could be considered very risky. On the other hand, for a product in which there is some degree of expected variability, the proportion might not be a practical concern for worry, in spite of the statistical significance.

## 5.5 Tests for Significant Similarity

Another approach to the problem of demonstrating product similarity or equivalence was presented by Meilgaard et al. (2006). It is based on the fact that we do not have to test against the chance performance level in applying the binomial test on proportions. Rather, we can test against some higher level of expected proportion correct, and see whether we are significantly *below* that level in order to make a decision about two products being sensorially similar enough. This is shown graphically in Fig. 5.2. Our usual tests for discrimination involve a statistical test against the chance performance level and we look for a point at which the confidence interval around our observed proportion no longer overlaps the chance level. This is just another way of thinking about the minimum level required for a significant difference. When the error bar no longer overlaps the chance level, we put that minimum number correct (for a given *N*) in our tables for the triangle test, duo–trio, etc. A higher proportion correct will be less likely to overlap and a larger sample size will shrink the error bars. As the "*N*" increases, the standard error of the proportion gets smaller. Thus higher proportions correct and larger panel sizes lead to less likely overlap with the chance level and thus a significant difference.

However, we can also test to see whether we are below some other level. The binomial test on proportions can be applied to other benchmarks as well. How can we determine this other benchmark? Once again, our proportion of allowable discriminators will give us a value to test against. We may have a very conservative situation, in which we can allow no more than 10% discriminators, or we might have a less critical or discerning public and be able to allow 30% or 50% discriminators or more. From the proportion of discriminators, it becomes a simple matter to calculate the other proportion we should test against, and see

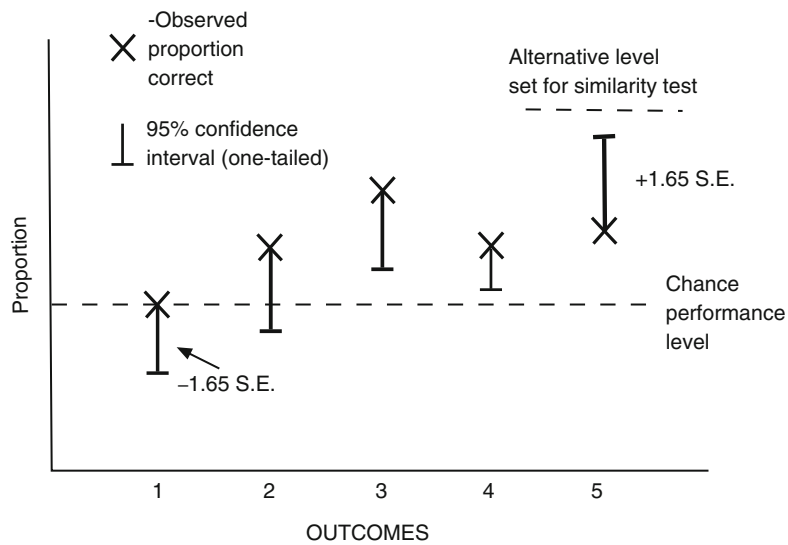



**Fig. 5.2** Difference testing and similarity testing outcomes. In outcome one, the performance is at the chance level, so there is obviously no difference. In the outcome, two performances are above chance, but the 95% one-tailed confidence interval overlaps the chance level, so no statistically significant difference is found. This level would be below the tabulated critical number of correct answers for that number of judges. In outcome three, the level correct and the associated confidence level are now above the chance level so a statistically significant difference is found. In outcome four, the level correct is lower than the third example, but the standard error has become smaller, due to an increase of *N*, so the outcome is also significant. In the fifth example, there is a significantly significant similarity, because the outcome and its associated one-tailed confidence interval are *below* the acceptable level based on the maximum allowable proportion of discriminators.

whether we are below that level. This is simply using Abbott's formula in reverse (Eq. (5.3)).

Tables H1 and H2 show critical values for significant similarity in the triangle test and for the duo–trio. Other tables are given in Meilgaard et al (2006). Here is a worked example. Suppose we conduct a triangle test with 72 panelists and 28 choose the correct (odd) sample. Do we have evidence for significant similarity? From Table H1 we see that for a criterion of no more than 30% discriminators, the critical value for a beta-risk of 10% is 32. Because we are *below* this value, we can conclude that the products are significantly similar.

If you examine these tables closely, you will note that there is a very narrow window for some proportions of discriminators and for a low number of panelists. There are empty cells in the table since we need a large panel size and low standard errors (small confidence intervals) in order to squeeze our result and the confidence interval between our test proportion and the chance proportion. The chance proportion forms, of course, a lower bound because it is not expected to see performance below chance. Once again, as in our power calculations, having a large sample size may be necessary to protect against Type II error, and probably a larger number of judges than we employ in most difference testing.

Let us look at this approach in some detail. The method of similarity testing is based upon the comparison of a maximum allowable proportion of discriminators to a confidence interval around the observed proportion correct. You can think of this test as involving three steps. First, we set an upper limit on the proportion of acceptable discriminators. Note that this involves professional judgment, knowledge of the products, and the business situation regarding consumer expectations about product consistency. It is not found in any statistics book. This is the same process as we discussed in Section 5.3 for choosing the size of the difference we need to detect. Second, we translate this into an expected percent correct by working through Eq. (5.3). The test then compares the inferred proportion plus its upper confidence interval to the maximum allowable proportion you set first. If the calculated value is less than the acceptable limit, we can conclude statistically significant similarity.

Here is the derivation of the formula. There are two items we need to know, the proportion correct we would expect, based on our proportion of discriminators, and the confidence interval boundaries around observed proportion correct given the number of judges. The proportion of discriminators and proportion expected correct are calculated just as in Eq. (5.3). The confidence interval of a proportion is given by $\pm$ $Z$ (standard error), where $Z$ = normal deviate for our desired level of confidence. For an upper one-tailed confidence interval at 95%, $Z = 1.65$. Equation (5.5) shows the standard error of the proportion, $SE_p$ and Eq. (5.6) the confidence interval:

$$SE_p = \sqrt{\frac{(X/N)(1 - X/N)}{N}} = \sqrt{pq/N} \tag{5.5}$$

where $X$ is the number correct, $N$ is the total number of judges, $p = X/N$ and $q = 1–p$.

$$CI_{95\%} = (X/N) \pm Z(SE_p) \tag{5.6}$$

where $Z$ is the $Z$-score for 95% confidence. The remaining step is to recast our confidence interval to include the fact that we are working with a limit on the number of discriminators, not the simple percent correct. For the triangle test, for example, the proportion of discriminators, $D/N$, is $1.5(X/N)–0.5$. Note that the standard error associated with the proportion of discriminators also is 1.5 times as large as the standard error associated with our observed proportion. So, our upper confidence interval boundary for discriminators now becomes

$$\text{Upper } CI_{95\%} = [1.5(X/N) - 0.5] + 1.645(1.5)\sqrt{\frac{(X/N)(1-X/N)}{N}} \tag{5.7}$$

Here is a worked example. Suppose we do a triangle test with 60 panelists, and 30 get the correct answer. We can ask the following questions: What is the best estimate of the number of discriminators? The proportion of discriminators? What is the upper 95% confidence interval on the number of discriminators? What is the confidence interval on the proportion of discriminators? Finally, could we conclude that there is significant similarity, based on a maximum allowable proportion of discriminators of 50%?

The solution is as follows: Let $X$ = number correct, $D$ = number of discriminators, so $X = 30$ and $N = 60$. We have $1.5(30)–0.5(60) = D$ or 15 discriminators, or 25% of our judges detecting the difference, as our best estimate.

The standard error is given by

$$1.5\sqrt{\frac{(30/60)[1-(30/60)]}{60}} = 0.097 = 9.7\%$$

and the upper bound on the confidence interval is given by Eq. (5.7), or $z$(SE) + proportion of discriminators = 1.65 (0.097) + 0.25 = 0.41 (or 41%).

So if our maximum allowable proportion of discriminators was 50%, we would have evidence that 95% of the time we would fall below this acceptable level. In fact, we would have 41% discriminating or less, given our observed percent correct of 50% which gives us our calculated best estimate of discriminators at 25%. This worked example is given to illustrate the approach. For practical purposes, the tables shown in Meilgaard et al. (2006) can be used as a simple reference without performing these calculations.

A similar approach was taken by Schlich (1993). He combined the questions of difference and similarity by calculating simultaneous alpha- and beta-risks for different panel sizes at the critical number correct for significant differences. Some of these values are shown in Table N. The table has two entries, one for the required number of judges and a second for the critical number correct at the crossover point. If the observed number correct is equal to or higher than the tabulated value, you can conclude that there is a significant difference. If the number correct in your test is lower, you can conclude significant similarity, based on the allowable proportion of discriminators you have chosen as an upper limit and the beta-risk. These tables can be very useful, but they required that you adopt the specified panel size that is stipulated for the conditions you have chosen for beta and proportion of discriminators.

## 5.6 The Two One-Sided Test Approach (TOST) and Interval Testing

The notion that equivalence can be concluded from a non-significant test, even with high test power, has largely been rejected by the scientific community concerned with bioequivalence (Bi, 2005, 2007). For example, the FDA has published guidelines for bioequivalence testing based on an interval testing approach (USFDA, 2001). This kind of test demands that the value of interest falls inside some interval and thus is sometimes referred to as interval testing. In general, this kind of testing is done on some scaled variable, like the amount of a drug delivered to the bloodstream in a certain specific period. Such a scaled variable is different from most discrimination tests, which are based on proportions, not some measured quantity that varies continuously. However, some scaled sensory data may fall under this umbrella, such as descriptive data or consumer acceptability ratings on a hedonic scale. Discrimination tests and preference tests are also amenable to this approach (Bi, 2006a; MacRae, 1995).

Logically, an interval test can be broken down into two parts, one test against an upper acceptable limit and one test against a lower acceptable limit. This is similar to finding some confidence intervals for an acceptable range of variation. In the case of many discrimination tests only the upper limit is of interest. The situation can then be a single one-tailed test. For paired comparison tests, the TOST method is described in detail by Bi (2007). In this article he shows some differences between the TOST estimates and the conventional two-sided confidence interval approach. Some authors recommend comparing the interval testing approach at 100(1–$\alpha$/2) to TOST because the interval testing approach at 100(1–$\alpha$) is too conservative and lacks statistical power (Gacula et al., 2009).

For scaled data, we may wish to "prove" that our test product and control product have mean values within some acceptable range. This approach can be taken with descriptive data, for instance, or acceptability ratings or overall degree-of-difference ratings. Bi (2005) described a non-central $t$-test for this situation. This is similar to a combination of $t$-tests in which we are testing that the observed difference between the means falls within some acceptable limit. An example of this approach is shown in Appendix at the end of this chapter. For purposes of using these models for equivalence, the sensory professional is advised to work closely with a statistical consultant. Further information on formal equivalence testing can be found in Welleck (2003) and Gacula et al. (2009).

Alternatives to the TOST method have been given by Ennis and Ennis (2010) and have some advantages to TOST from a statistical perspective. An alternative to TOST that is applicable to a non-directional 2-AFC (e.g., a two-tailed 2-AFC, much like a paired preference) has been proposed (Ennis and Ennis, 2010). Note that under this approach, establishing an equivalence or

parity situation usually requires a much larger sample size ($N$) than simple tests for difference. Useful tables derived from this theory are given in Ennis (2008). The theory states that a probability value for the equivalence test can be obtained from an exact binomial or more simply from a normal approximation as follows:

$$p = \phi\left(\frac{|x| - \theta}{\sigma}\right) - \phi\left(\frac{|-x| + \theta}{\sigma}\right) \tag{5.8}$$

where phi ($\Phi$) symbolizes the cumulative normal distribution area (converting a $z$-score back to a $p$-value), theta ($\theta$) is the half-interval for parity such as $\pm 0.05$, and sigma ($\sigma$) is the estimated standard error of the proportion (square root of $pq/N$). $x$ in this case is the difference between the observed proportion and the null (for 2-AFC, subtract 0.5). A worked example is given in Chapter 13, Section 13.3.5. Note that the tables given in Ennis (2008) are specific to the 2-AFC and may not be used for other tests such as the duo–trio.

## 5.7 Claim Substantiation

A special case of equivalence testing arises when a food or consumer product manufacturer wishes to make a claim of equivalence or parity against a competitor. Such a claim might take the form of a statement such as "our product is just as sweet as product X." Because of the legal ramifications of this kind of test, and the need to prove that the result lies within certain limits, large numbers of consumers are typically required for such a test, with recommended sample sizes from 300 to 400 as a minimum (ASTM, 2008a). This is a different scenario than most simple discrimination tests that use laboratory panels of 50–75 judges. The special case of proving preference equality (products chosen equally often in a preference test) is discussed further in Chapter 19 on strategic research.

The simple case of a paired comparison test (2-AFC) is amenable to this kind of analysis. As noted above, Bi (2007) discussed the TOST approach to equivalence for 2-AFC with worked examples. There are two different statistical scenarios: In one case we wish to make an equality claim, and in the second, we want to make a claim that our product is "unsurpassed." The equality claim involves two tests, because neither product can have more of the stated attribute than another. The unsurpassed claim is a simple one-tailed alternative and just requires showing that our product is not significantly lower than the other product. For both of these tests, we have to choose some lower bound that we cannot cross. For example, a common criterion for the equality claim is that the true population percentage in the paired test lies between 45 and 55% of choices. Thus a 5% difference is considered "close enough" or of no appreciable practical significance.

The equality claim requires that neither one of the products cross over the lower bound and can be viewed as two one-tailed tests. Tables for the minimum number allowable in such tests can be found in ASTM (2008a). For the unsurpassed claim we are stating that our product is not inferior or not lower in the attribute in question. For this purpose the test takes the following form of a simple binomial-approximation $Z$-score:

$$Z = \frac{(P_{\text{obs}} - 0.45) - (1/2N)}{\sqrt{\frac{(0.45)(0.55)}{N}}} \tag{5.9}$$

where $P_{\text{obs}}$ is the proportion observed for your test product. In the case of large sample sizes ($N$>200), the value of the continuity correction, $1/2N$, becomes negligible. Note that this is a one-tailed test, so the obtained $Z$ must be greater than or equal to 1.645. If the obtained $Z$ is greater than that value, you have support for a claim that your product is not lower than the competitor. In the case of our sweetness claim, we can be justified in saying our product is "just as sweet."

## 5.8 Models for Discrimination: Signal Detection Theory

In the preceding sections, we looked at the size of the sensory difference in terms of the proportion of panelists who could be discriminating in any given test. The calculations are based on an adjustment for chance correct guesses. However, the chance probability level does not tell the whole story, because some tests are harder or involve more variability than others, even at the same chance probability level. As an example, the triangle test is "more difficult" than the 3-AFC test, because it is a more difficult cognitive task to find

the odd sample (one that entails higher variability), as opposed to simply choosing the strongest or weakest. In this section, we will look at a more sophisticated model in which this issue can be taken into account. From this theory, we can derive a universal index of sensory similarity or difference and one that takes into account the difficulty or variability inherent in different discrimination tests.

One of the most influential theories in psychophysics and experimental psychology is signal detection theory (SDT). The approach grew from the efforts of engineers and psychologists who were concerned with the decisions of human observers under conditions that were less than perfect (Green and Swets, 1966). An example is the detection of a weak signal on a radar screen, when there is additional visual "noise" present in the background. Mathematically, this theory is closely related to an earlier body of theory developed by Thurstone (1927). Although they worked in different experimental paradigms, credit should be given to Thurstone for the original insight that a scaled difference could be measured based on performance and error variability. Although signal detection usually deals with threshold-level sensations, any question of perceived differences (when such differences are small) can be addressed by SDT. For a good introduction to SDT, the book on psychophysics by Baird and Noma (1978) is useful and for a more detailed look, Macmillan and Creelman (1991) is recommended.

### 5.8.1 The Problem

In traditional threshold experiments, the physical strength of a weak stimulus is raised until the level is found where people change their responses from "No, I do not taste (or smell, see, hear, feel) anything" to "Yes, now I do." This is the original procedure for the method of limits (see Chapter 2). The difficulty in such an experiment is that there are many biases and influences that can affect a person's response, in addition to the actual sensation they experience. For example, they may expect a change in sensation and anticipate the level where something will be noticeable. Conversely, a person might adopt a very conservative stance and want to be very sure that something is clearly perceivable before they respond. An example in industry might be in quality control, where the mistaken rejection of a batch of product could incur a large cost if the batch has to be reworked or discarded. On the other hand, a high-margin upscale product with a brand-loyal and knowledgeable consumer base might require narrower or more stringent criteria for product acceptance. Any sensory problem at all might cause rejection or retesting of a batch. The criterion for rejection would cast a wider net to be sure to catch potential problem items before they can offend the brand-loyal purchasers.

So a decision process is layered on top of the actual sensory experience. A person may set a criterion that is either conservative or lax in terms of how much evidence they need to respond positively. Here is a simple example of a decision process involved in perception: Suppose you have just purchased a new pair of stereo headphones. It is the end of a long work week and you have gone home to enjoy your favorite music in a comfortable chair. As you settle in and turn the music up very loud, you think the phone might be ringing. Let us consider two scenarios: In one case, you are expecting a job offer following a successful interview. Do you get up and check the phone? In another case you are expecting a relative to call who wants to borrow money. Do you get up? What are the relative payoffs and penalties associated with deciding that the phone has actually rung? What does it take to get you up out of that comfortable chair?

### 5.8.2 Experimental Setup

In a classic signal detection experiment, two levels of a stimulus are to be evaluated, for example, the background or blank stimulus called the "noise" and some weak but higher level of stimulus intensity near threshold called the "signal" (Baird and Noma, 1978; Macmillan and Creelman, 1991) Stimuli are presented one at a time and the observer would have to respond "Yes, I think that is a case of the signal," or "No, I think that is the case of the noise." So far this resembles the A, not-A test in sensory evaluation. Both signal and noise would be presented many times and the data would be tabulated in a two-by-two response matrix as shown in Fig. 5.3. Over many presentations, some correct decisions would be made when a signal is in

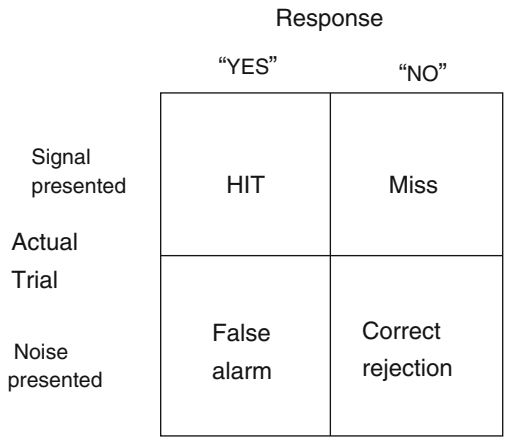



**Fig. 5.3** The stimulus–response matrix for a signal detection experiment. The hit rate is the proportion of times the subject responds "yes" or "signal" when in fact the signal was presented. The false alarm rate is the proportion of noise trials when the subject also responds "yes" or "signal" (noise presented). These two outcomes define the response space since the misses are the total of the signal trials (which the experimenter has designed into the study) minus hits, and the correct rejections are likewise the number of noise trials minus false alarms (there is only one degree of freedom per row).

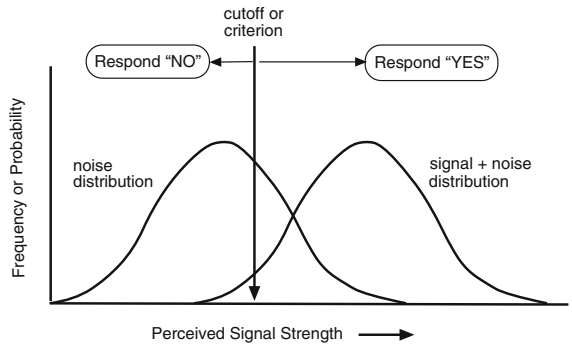



**Fig. 5.4** Signal detection assumptions include normally distributed experiences from signal and noise trials, with equal variance and the establishment of a stable criterion or cutoff level, above which the subject responds "yes" and below which the subject responds "no".

fact presented and these are called "hits" in signal detection terminology. Since the stimuli are confusable, sometimes the observer would respond positively on noise trials as well, mislabeling them as signal. These are called "false alarms." There are also situations in which the signal is presented and the observer fails to call it a signal (a "miss") and cases in which the noise trial is correctly labeled a noise trial. However, since we have set up the experimental design and know how many signal and noise trials we have presented, the total number of signal trials is equal to the hits plus misses and the total number of noise trials equals false alarms plus correct rejections. In other words, there is only one degree of freedom in each row and we can define the observer's behavior by examining only hit and false alarm rates.

### 5.8.3 Assumptions and Theory

The theory makes a few sensible assumptions (Baird and Noma, 1978; Green and Swets, 1966). Over many trials, the sensations from signal and noise are normally distributed with equal variance. That is, sometimes a more intense sensation will be felt from the signal, and sometimes a weaker sensation, and over many trials these experiences will be normally distributed around some mean. Similarly the noise trials will sometimes be perceived as strong enough so that they are mistakenly called a "signal." Once the observer is familiar with the level of sensations evoked, he or she will put a stable criterion in place. When the panelist decides that the sensation is stronger than a certain level, the response will be "signal" and if weaker than a certain amount a "noise" response will be given. This situation is shown in Fig. 5.4. Remember that the panelist does not know if it is a signal or noise trial, they just respond based on how strong the sensation is to them.

Variabilities in the signal and noise are reasonable assumptions. There is spontaneous variation in the background level of activity in sensory nerves, as well as variance associated with the adaptation state of the observer, variation in the stimulus itself, and perhaps in the testing environment. The greater the overlap in the signal and noise distributions, the more difficult the two stimuli are to tell apart. This shows up in the data as more false alarms relative to the occurrence of hits. Of course, in some situations, the observer will be very cautious and require a very strong sensation before responding "yes, signal." This will not only minimize the false alarm rate but also lower the hit rate. In other situations, the observer might be very lax, and respond "yes" a lot, producing a lot of hits, but at the cost of increased false alarms. The hit and false alarm rates will co-vary as the criterion changes.

Now we need to connect the performance of the observer (Fig. 5.3) to the underlying scheme in Fig. 5.4 to come up with a scaled estimate of performance and one that is independent of where observers set their particular criteria for responding. The separation of the two distributions can be specified as the difference between their means and the unit of measurement as the standard deviations of the distributions (a convenient yardstick). Here is the key idea: The proportion of signal trials that are hits corresponds to the area underneath the signal distribution to the right of the criterion, i.e., the sensation stronger than our cutoff, so response is "yes" to a signal presentation. Similarly, the proportion of false alarm trials represents the tail of the noise distribution to the right of the cutoff, i.e., sensations stronger than criterion but drawn from the noise experiences. This scheme is shown in Fig. 5.5.

All we need to estimate, then, is the distance from the criterion level or cutoff to the mean of each distribution. These can be found from the *z*-scores relating the proportion to the distance in standard deviation units. Since we know the relationship between proportions and *z*-scores, the two distances

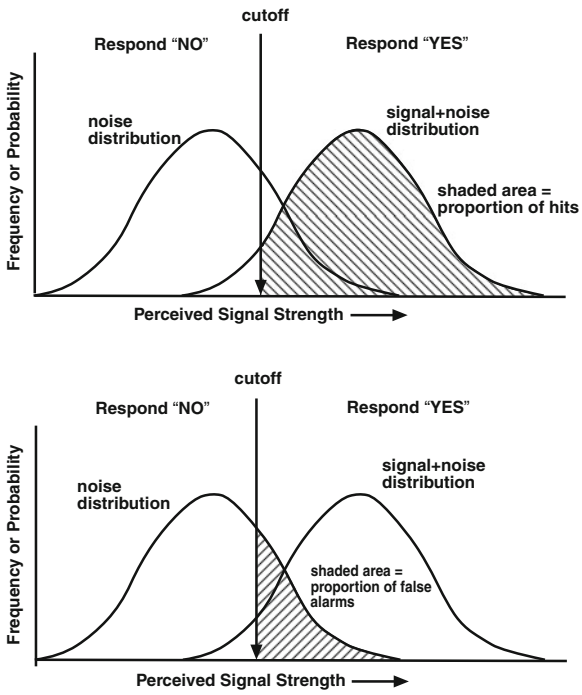



**Fig. 5.5** Signal and noise distributions are shaded to show the area corresponding to the proportions of hits and false alarms, respectively. These proportions can then be converted to *z*-scores.

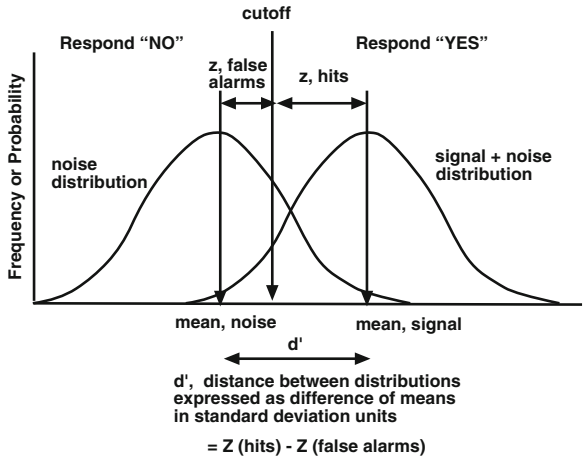



**Fig. 5.6** How $d'$ is calculated based on the signal detection scheme. Using the conversion of proportions (areas) to *z*-scores, the overall difference ($d'$, pronounced "d prime") is given by the *z*-value for hits minus the *z*-value for false alarms.

can be estimated, and then summed, as shown in Fig. 5.6. Due to the way that *z*-scores are usually tabulated, this turns out to be a process of subtraction and the value of sensory difference called d′ ("*d*-prime") is equal to the *z*-score for the proportion of hits minus the *z*-score for the proportion of false alarms.

### 5.8.4 An Example

The great advantage of this approach is that we can estimate this sensory difference independently of where the observer sets the criterion for responding. Whether the criterion is very lax or very conservative, the hit and false alarm *z*-scores will change to keep $d'$ the same. The criterion can slide around, but for a given set of products for the same panelist, the difference between the two distributions remains the same. When the criterion is moved to the right, fewer false alarms result and also fewer hits. (Note that the *z*-score will change sign when the criterion passes the mean of the signal distribution). If the criterion becomes very lax, the hit and false alarm rates will both go up and the *Z*-score for false alarms will change sign if the proportion of false alarms is over 50% of the noise trials. Table 5.2 may be used for conversion of proportions of hits and false alarms to *z*-scores. Figure 5.7 shows a criterion shift for two approximately equal levels of discriminability. The upper panel shows a conservative criterion with only 22% hits and 5% false

**Table 5.2** Proportions and *Z*-scores for calculation of $d'$

| Proportion | *Z*-score | Proportion | *Z*-score | Proportion | *Z*-score | Proportion | *Z*-score |
|---|---|---|---|---|---|---|---|
| 0.01 | −2.33 | 0.26 | −0.64 | 0.51 | 0.03 | 0.76 | 0.71 |
| 0.02 | −2.05 | 0.27 | −0.61 | 0.52 | 0.05 | 0.77 | 0.74 |
| 0.03 | −1.88 | 0.28 | −0.58 | 0.53 | 0.08 | 0.78 | 0.77 |
| 0.04 | −1.75 | 0.29 | −0.55 | 0.54 | 0.10 | 0.79 | 0.81 |
| 0.05 | −1.64 | 0.30 | −0.52 | 0.55 | 0.13 | 0.80 | 0.84 |
| 0.06 | −1.55 | 0.31 | −0.50 | 0.56 | 0.15 | 0.81 | 0.88 |
| 0.07 | −1.48 | 0.32 | −0.47 | 0.57 | 0.18 | 0.82 | 0.92 |
| 0.08 | −1.41 | 0.33 | −0.44 | 0.58 | 0.20 | 0.83 | 0.95 |
| 0.09 | −1.34 | 0.34 | −0.41 | 0.59 | 0.23 | 0.84 | 0.99 |
| 0.10 | −1.28 | 0.35 | −0.39 | 0.60 | 0.25 | 0.85 | 1.04 |
| 0.11 | −1.23 | 0.36 | −0.36 | 0.61 | 0.28 | 0.86 | 1.08 |
| 0.12 | −1.18 | 0.37 | −0.33 | 0.62 | 0.31 | 0.87 | 1.13 |
| 0.13 | −1.13 | 0.38 | −0.31 | 0.63 | 0.33 | 0.88 | 1.18 |
| 0.14 | −1.08 | 0.39 | −0.28 | 0.64 | 0.36 | 0.89 | 1.23 |
| 0.15 | −1.04 | 0.40 | −0.25 | 0.65 | 0.39 | 0.90 | 1.28 |
| 0.16 | −0.99 | 0.41 | −0.23 | 0.66 | 0.41 | 0.91 | 1.34 |
| 0.17 | −0.95 | 0.42 | −0.20 | 0.67 | 0.44 | 0.92 | 1.41 |
| 0.18 | −0.92 | 0.43 | −0.18 | 0.68 | 0.47 | 0.03 | 1.48 |
| 0.19 | −0.88 | 0.44 | −0.15 | 0.69 | 0.50 | 0.94 | 1.55 |
| 0.20 | −0.84 | 0.45 | −0.13 | 0.70 | 0.52 | 0.95 | 1.64 |
| 0.21 | −0.81 | 0.46 | −0.10 | 0.71 | 0.55 | 0.96 | 1.75 |
| 0.22 | −0.77 | 0.47 | −0.08 | 0.72 | 0.58 | 0.97 | 1.88 |
| 0.23 | −0.74 | 0.48 | −0.05 | 0.73 | 0.61 | 0.98 | 2.05 |
| 0.24 | −0.71 | 0.49 | −0.03 | 0.74 | 0.64 | 0.99 | 2.33 |
| 0.25 | −0.67 | 0.50 | 0.00 | 0.75 | 0.67 | 0.995 | 2.58 |

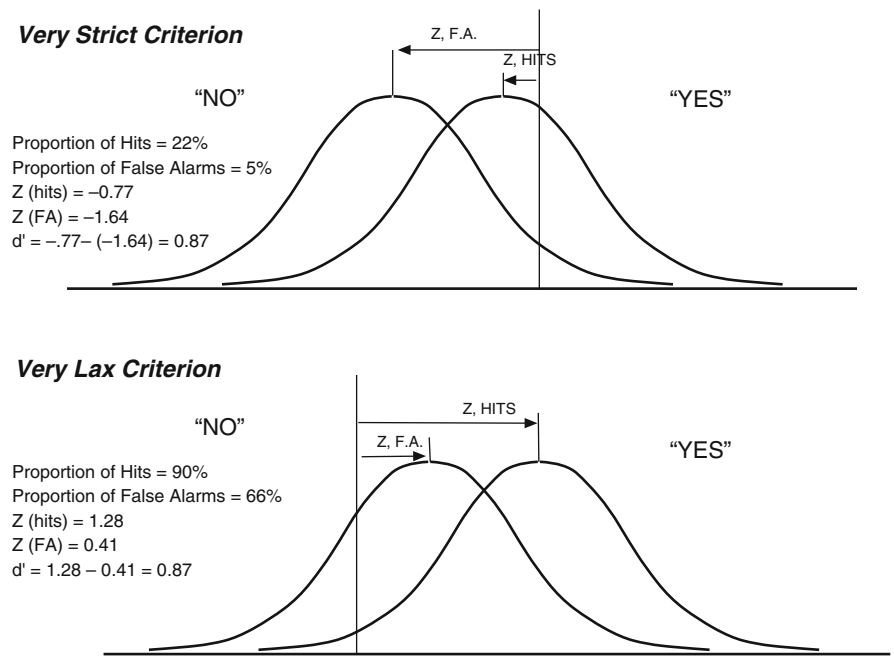



**Fig. 5.7** The measure of sensory difference, $d'$, will remain constant for the same observer and the same stimuli, even though the criterion shifts. In the upper case, the criterion is very strict and the observer needs to be sure before responding "yes" so there is a low proportion of hits but also a low proportion of false alarms. In the lower panel the subject sets a lax criterion, responding "yes" most of the time, catching a lot of hits, but at the expense of many false alarms.

alarms. Referring to Table 5.2, the *Z*-scores for these proportions are –0.77 and –1.64, respectively, giving a $d'$ value of –0.77–(–1.64) or +0.87. The lower panel illustrates a less conservative mode of responding, with 90% hits and 66% false alarms. Table 5.2 shows the *Z*-scores to be 1.28 and 0.41, again giving a $d'$ of 0.87.

In other words, the $d'$ does not change, even though the criterion has shifted. This theory permits a determination of the degree of sensory discriminability, regardless of the bias or criterion of the observer. In the next section, we will examine how the theory can be extended to include just about any discrimination test.

### 5.8.5 A Connection to Paired Comparisons Results Through the ROC Curve

How can we connect the SDT approach to the kinds of discrimination tests used in sensory evaluation? One way to see the connection is to look at the receiver operating characteristic or ROC curve. This curve defines a person's detection ability across different settings of the criterion. In the ROC curve, hit rate in different situations is plotted as a function of false alarm rate. Figure 5.8 shows how two observers would behave in several experiments with the same levels of the stimulus and noise. Payoffs and penalties could be varied to produce more conservative or more lax behaviors, as they often were in the early psychophysical studies. As criterion shifts, the performance moves along the characteristic curve for that observer and for those particular stimuli. If the hit rate and false alarm rates were equal, there is no discrimination of the two levels, and $d'$ is zero. This is shown by the dotted diagonal line in the figure. Higher levels of discrimination (higher levels of $d'$) are shown by curves that bow more toward the upper left of the figure. Observer 2 has a higher level of discrimination, since there are more hits at any given false alarm rate or fewer false alarms at a given hit rate. The level of discrimination in this figure, then is proportional to the area under the ROC curve (to the right and below), a measure that is related to $d'$. Note that the dotted diagonal cuts off one-half of the area of the figure. One-half (50%) is the performance you would expect in a paired comparison test if there were no difference between the products. From this you can see that there should be a correspondence between the area under the ROC curve (which is proportional to $d'$) and the performance we would expect in a 2-AFC or paired comparison test. Results from other kinds of discrimination tests such as the triangle, duo–trio, and 3-AFC can be mathematically converted to $d'$ measures (Ennis, 1993).

## 5.9 Thurstonian Scaling

Thurstone was dealing with the kinds of studies done in traditional psychophysics, like the method of constant stimuli (see Chapter 2). This method is basically just a series of paired comparisons against a constant or standard stimulus. Thurstone realized that if you got 95% correct in a paired test, the sensory difference ought to be bigger than if you only got 55% correct. So he set out to come up with a method to measure the degree of difference, working from the percent correct in a paired test. In doing this he formulated a law of comparative judgment (Thurstone, 1927).

### 5.9.1 The Theory and Formulae

Thurstone's law of comparative judgment can be paraphrased as follows: Let us assume that the panelist will compare two similar products, A and B, over several trials and we will record the number of times A is judged stronger than B. Thurstone proposed that the sensory events produced by A and B would be normally distributed. Thurstone called these perceptual distributions "discriminal dispersions," but they are analogous to the distributions of signal and noise events in the yes/no task. The proportion of times A is judged stronger than B (the datum) comes from a process of mental comparison of the difference between the two stimuli. Evaluating a difference is analogous to a process of subtraction. Sometimes the difference will be positive (A stronger than B) and sometimes the difference will be negative (B stronger than A) since the two items are confusable. One remaining question is how the sampling distribution for these differences would arise from the two underlying discriminal dispersions as shown in Fig. 5.9. The laws of statistics can help here, since it is possible to relate the difference sampling distribution (lower curve) to

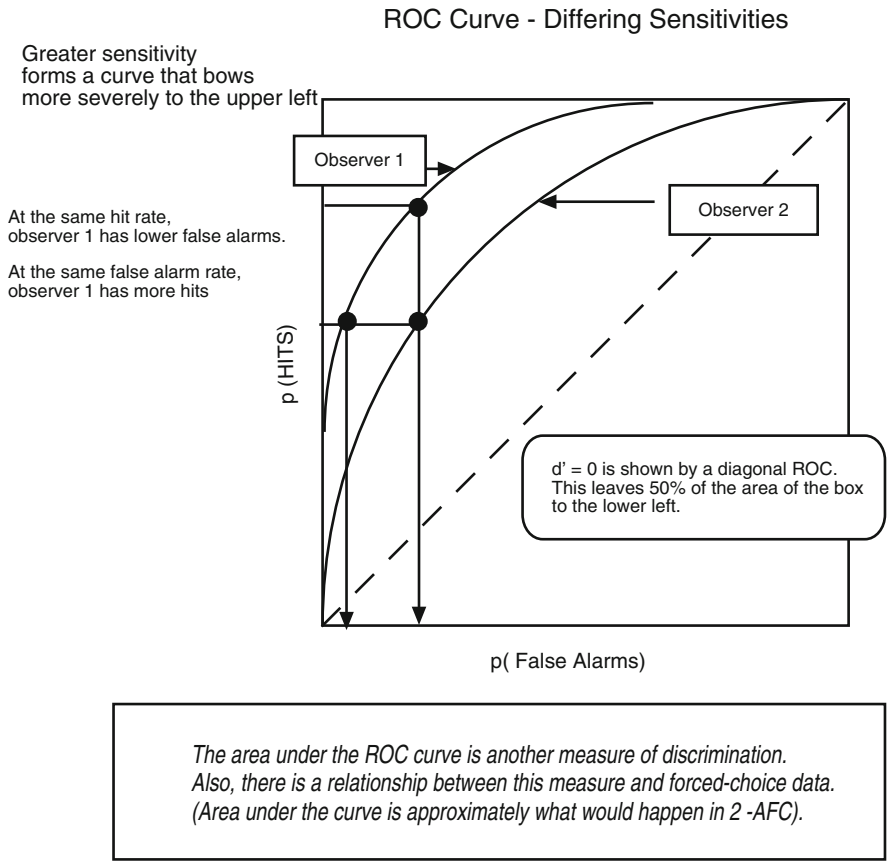



**Fig. 5.8** The ROC curve, or receiver operating characteristic, shows the behavior of a single individual under various criterion shifts, plotting the proportion of hits against the proportion of false alarms. Better discrimination (higher $d'$) is shown by a curve that bows more toward the upper left corner. Thus observer 1 has better performance and better discrimination ability than observer 2. The area under the ROC curve (to the right and below) is another measure of discrimination and can be converted to $d'$ values. Note that the diagonal describes no discrimination, when hits always equal the rate of false alarms. Also note that the area of the box below and to the right is 50% which would be the performance in a paired test or 2-AFC when there is no difference. Thus the area under the ROC curve is expected to be proportional to the performance one would observe with a given observer and a given pair of stimuli in a 2-AFC test.

the sensory events depicted by the upper distributions. This statistical result is given by the following equation:

$$S_{\text{diff}} = \sqrt{S_{\text{a}}^2 + S_{\text{b}}^2 + 2rS_{\text{a}}S_{\text{b}}} \tag{5.10}$$

$$M_{\text{difference}} = Z\sqrt{S_{\text{a}}^2 + S_{\text{b}}^2 + 2rS_{\text{a}}S_{\text{b}}} \tag{5.11}$$

where $M$ is the difference scale value, $Z$ is the $z$-score corresponding to the proportion of times A is judged stronger than B, and $S_{\text{a}}$ and $S_{\text{b}}$ are the standard deviations of the original discriminal dispersions. The "$r$" represents the correlation between sensations from A and B, which might be negative in the case of contrast or positive in the case of assimilation. If we make the assumptions that $S_{\text{a}}$ and $S_{\text{b}}$ are equal and $r = 0$ (no sequential dependency of the two stimuli) then the equation simplifies to

$$M = Z\sqrt{2}S \tag{5.12}$$

where $S$ is the common standard deviation. These simplified assumptions are referred to as "Thurstone's Case V" in the statistical literature (Baird and Noma, 1978). The mean of the difference scores is statistically the same as the difference of the two means. So to get to $d'$, which is the mean difference divided by the original standard deviation, we have to multiply our $z$-score (from the percent correct) by the square

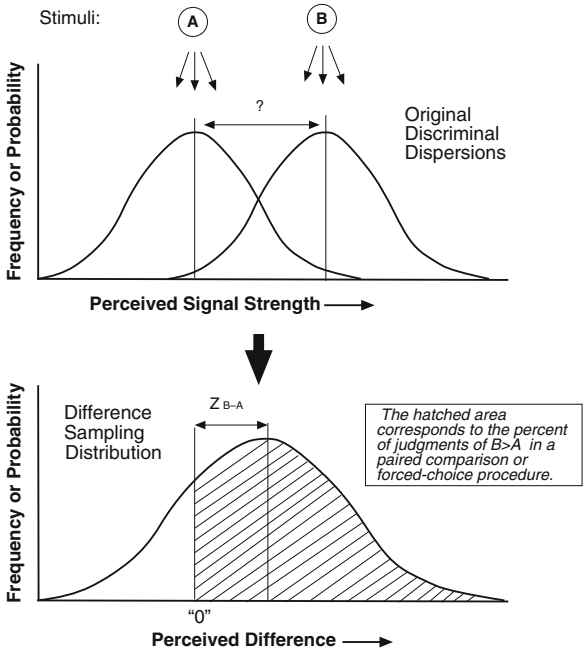



**Fig. 5.9** The Thurstone model proposes that the proportion of times one stimulus is judged greater than another is predicted by a difference sampling distribution, which in turn arises from the sensory variability and degree of overlap in the original stimuli.

root of 2. In other words, the *z*-score value is smaller than what would be estimated from the $d'$ of the yes/no signal detection experiment by the square root of two (Macmillan and Creelman, 1991). The distance of the mean from an arbitrary zero point can be determined by a *z*-score transformation of the proportion of times A is judged stronger than B. We can conveniently work with the zero point as the mean of distribution for the weaker of the two stimuli (Baird and Noma, 1978). Like $d'$, this gives us a measure that is expressed in standard deviation units.

### 5.9.2 Extending Thurstone's Model to Other Choice Tests

We can extend this kind of scale value to any kind of choice test and tables have been published for various conversions of percent correct to $d'$ or delta values (Bi, 2006a; Ennis, 1993; Frijters et al., 1980; Ura, 1960). Delta is sometimes used to refer to a population variable (rather than $d'$ which is a sample statistic) but the meaning is the same in terms of the sensory difference it describes. Other theorists saw the applicability of a signal detection model to forced-choice data. Ura (1960) and Frijters et al. (1980) published the mathematical relationships to relate triangle test performance to $d'$ as well as other test procedures commonly used in food science.

Ennis (1990, 1993) has examined Thurstonian models as one way of showing the relative power of difference tests. Like Frijters, he showed that for a given level of scaled difference a lower level of percent correct is expected in the triangle test, as opposed to the 3-AFC test. On the basis of the variability in judging differences versus intensities, one expects higher performance in the 3-AFC test. This has become famous as the "paradox of the discriminatory non-discriminators" (Byer and Abrams, 1953; Frijters, 1979). In the original paper, Byer and Abrams noted that it was possible for many panelists to answer incorrectly under the triangle test instructions, but still accurately choose the strongest sample when it was the odd one (or the weakest when it was the odd one). An example of how this could occur is shown in Fig. 5.10. That is, for the 3-AFC instructions, there was a higher percent correct. Frijters (1979) was able to show that the different percents correct for the triangle and 3-AFC in Byer and Abram's data actually yielded the same $d'$ value, hence resolving the apparent

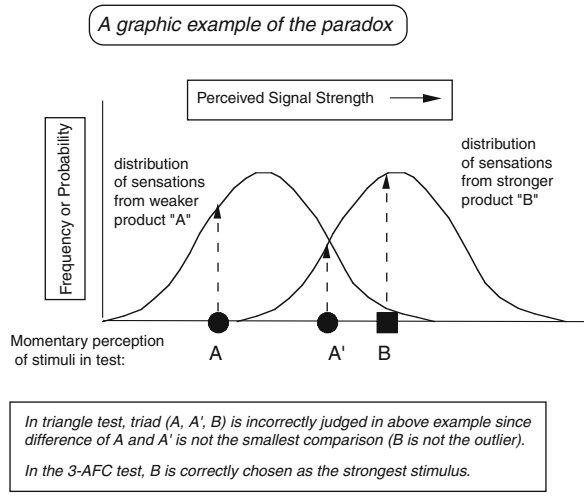



**Fig. 5.10** An example of the paradox of the discriminatory non-discriminators. In a hypothetical triad of products, the odd sample, B, is chosen as the strongest sample, a correct answer for the 3-AFC test. However, one of the duplicate samples, A, is momentarily perceived as the outlier, leading to an incorrect choice as the odd sample under the triangle test instructions.

paradox. Delwiche and O'Mahony (1996) demonstrated that similar predictions can be made for tetradic (four-stimulus procedures).

The ability to reconcile the results from different experimental methods and convert them to a common scale is a major advantage of the signal detection and Thurstonian approaches (Ennis, 1993). Studies continue to show the constancy of $d'$ estimates across tasks and methods (Delwiche and O'Mahony, 1996; Stillman and Irwin, 1995), although there are occasional exceptions where other factors come into play to complicate the situation and require further model development (Antinone et al., 1994; Lawless and Schlegel, 1984). Applying the correct Thurstonian model requires that you understand the cognitive strategy of the panelist (O'Mahony et al., 1994). For example, am I looking for the smallest of three pairs of differences (a triangle strategy for finding the odd sample) or am I trying to discern the strongest of three intensities (a 3-AFC "skimming" strategy)? If one has a different strategy for a given test method or task, the resulting $d'$ value will not reflect what people are actually doing in the test. For example, if a number of panelists are "skimming" for the strongest sample, but have been given triangle instructions, the $d'$ will not make sense if taken from the triangle test tables.

Another complicating factor concerns sequential effects in groups of products that are presented at the same time. The discriminability of two items depends not only on the relative strength of signal versus noise sensations but also on the sequence in which items are presented. Thus a strong stimulus following a weak one (signal after noise trial) may give a stronger sensation of difference than a noise trial following a signal trial. O'Mahony and Odbert (1985) have shown how this leads to better performance for some discrimination tests over others in a theory called "sequential sensitivity analysis." Ennis and O'Mahony (1995) showed how sequential effects can be incorporated into a Thurstonian model. Another factor concerns the fact that most foods are multi-dimensional and the simple SDT and Thurstone models are usually formalized as a one-dimensional variation. Ennis and Mullen (1986) using a multivariate model showed how variation in irrelevant dimensions could degrade performance.

The important conclusion for sensory professionals to draw from this theory is that the common tests for overall difference, e.g., the triangle and duo–trio, are not very sensitive tests. That is, for a given $d'$ value, a much larger panel size needs to be tested to be sure that the difference is detected by the test. This is in comparison to the forced-choice procedures such as the paired comparison and 3-AFC tests which will detect a statistically significant difference for a given $d'$ at much smaller panel sizes (Ennis, 1993). Put a different way, for a given panel size, the triangle test could easily miss a difference that the 3-AFC test would detect, as seen in the Byer and Abrams "paradox." Unfortunately, when an ingredient or processing change is made in a complex food, one cannot always predict any simple singular attribute to use for the AFC tests, nor perhaps even overall strength of flavor, taste, etc. So the sensory professional is stuck using the less sensitive type of test. In the face of a statistically significant result, this is not a problem. But if the equivalence decision is based on a non-significant test outcome, the decision to conclude equivalence can be very risky.

## 5.10 Extensions of the Thurstonian Methods, *R*-Index

### *5.10.1 Short Cut Signal Detection Methods*

One additional method deserves mention in the application of signal detection models to the discrimination testing situation. A practical impediment to the application of signal detection theory to foods has been the large number of trials necessary in the traditional yes/no signal detection experiment. With foods, and especially in applied difference testing, it has rarely been possible to give the large numbers of trials to each subject needed to accurately estimate an individual's $d'$ value.

O'Mahony (1979) saw the theoretical advantage to signal detection measures, and proposed short-cut rating scale methods to facilitate the application of signal detection in food evaluations. The *R*-index is one example of an alternative measure developed to provide an index of discrimination ability but without the stringent assumptions entailed by $d'$, namely equal and normally distributed variances from signal and noise distributions. The area under the ROC curve is another measure of discrimination that does not depend upon

the exact forms of the signal and noise distributions (see Fig. 5.8). The *R*-index is one such measure and converts rating scale performance to an index related to the percentage of area under the ROC curve, a measure of discrimination. It also gives an indication of what we would expect as performance in two-alternative forced-choice task, which is of course, mathematically related to $d'$.

### 5.10.2 An Example

Here is an example of the *R*-index approach. In a typical experiment, 10 signal and 10 noise trials are given. Subjects are pre-familiarized with signal stimuli (called "A") and noise stimuli (called "B" in this example). Subjects are asked to assign a rating scale value to each presentation of A and B, using labels such as "A, definitely," "A, maybe", "B, maybe" and "B, definitely."

For a single subject, performance on the 20 trials might look like this:

| | Ratings | | | |
|---|---|---|---|---|
| | A, definitely | A, maybe | B, maybe | B, definitely |
| Signal presented | 5 | 2 | 2 | 1 |
| Noise presented | 1 | 2 | 3 | 4 |

Obviously, there is some overlap in these distributions and the stimuli are somewhat confusable.

*R* is calculated as follows: Pair every rating of the signal with every rating of the noise as if there were paired comparisons. In this example, there are $10 \times 10$ or 100 hypothetical pairings. *R* calculates how many times would the signal "A" be identified correctly or called the stronger of the pair. First, we consider the five cases in which signal (A) was rated "A, definitely." When paired against the $(2 + 3 + 4 =)$ 9 cases in which the noise trial (B) received a lower rating (i.e., noise was judged less like "A" than signal), 45 correct judgments would have been made if there were actually paired tests. For the five cases in which the signal was "A, definitely" were paired with noise rated "A, definitely," there are five ties $(5 \times 1)$, so we presume that half the trials (2.5) would be correct and half incorrect if a choice were forced. We then to continue to make these hypothetical pairings of each rating of A with ratings of B, based upon the frequencies in each cell of our matrix. Thus the ratings of signal as "A, maybe" give $2 \times 7 = 14$ "correct" pairings (i.e., A rated higher than B) and the ratings of signal as "B, maybe" give $2 \times 4$=8 correct pairings. There are 17 total ties (counted as 8.5 correct pairings). The *R*-index, then is 45 + 14 + 8 + 8.5 (for ties) = 75.5. In other words, our best guess about the total percent correct in a two-alternative forced-choice task like a paired comparison test would be about 75.5%.

This result indicates a slight degree of difference and that this pair of items is sometimes confusable. Obviously, as the two stimuli have less overlapping response patterns, there is better discrimination and a higher *R*-value. Remember that this would correspond to 75.5% of the area below the ROC curve for this person. Taking a *z*-score and multiplying by the square root of 2 gives us a $d'$ of 0.957 (Bi, 2006b). This value, close to one, also suggests that the difference is above threshold but not very clear. Statistical tests for the *R*-index, including confidence intervals for similarity testing are given in Bi (2006b).

As in other signal detection methods, the *R*-index allows us to separate discrimination performance from the bias or criterion a person sets for responding. For example, we might have an observer who is very conservative and calls all stimuli noise or labels them "B" in our example. If the observer assigned all A-trials (signals) to "B, maybe" and all B-trials (noise) to "B, definitely" then the *R*-index would equal 100, in keeping with perfect discrimination. The fact that all stimuli were considered examples of "B" or noise shows a strong response bias, but this does not matter. We have evidence for perfect discrimination due to the assignment of the two stimuli to different response classes, even though the observer was very biased to use only one part of the rating scale. Another advantage of *R*-index methods is that far fewer trials need be given as compared with the yes/no procedure.

## 5.11 Conclusions

A common issue in applied sensory testing is whether the experimental sample is close enough to a control or standard sample to justify a decision to substitute

it for the standard. This is an exceedingly difficult question from a scientific perspective, since it seems to depend upon proving a negative, or in statistical terms, on proving the null hypothesis. However, failure to reject the null can occur for many reasons. There may be truly no difference, there may be insufficient sample size ($N$ too low), or there may be too much variability or error obscuring the difference (standard deviations too high). Since this situation is so ambiguous, in experimental science we are usually justified in withholding a decision if we find no significant effect. Statisticians often say "there is insufficient evidence to reject the null" rather than "there is no significant difference." However, in industrial testing, the non-significant difference can be practically meaningful in helping us decide that a sample is like some control, as long as we have some knowledge of the sensitivity and power of our test. For example, if we know the track record of a given panel and test method for our particular product lines, we can sometimes make reasonable decisions based on a non-significant test result.

An alternative approach is to choose some acceptable interval of the degree of difference and see whether we are inside that interval or below some acceptable limit. This chapter has approached the degree-of-difference issue from two perspectives. The first was to convert our percent correct to an adjusted percent correct based on the traditional correction for guessing given by Abbott's formula. This allows us to estimate the percent of people actually discriminating, assuming a simple two-category model (either you see the difference or you guess). The second approach is to look at the degree of difference, or conversely the power of the test to detect that difference, as a function of a Thurstonian scaled value such as delta or $d'$. This value provides a more universal yardstick for sensory differences, as it takes into account the difficulty or variability inherent in the test, and also the cognitive strategy of the panelist in different tasks.

Note that there is an important limitation to the correction-for-guessing models. The guessing model and the Thurstonian model have different implications regarding the difficulty of the triangle and 3-AFC test. The guessing model only considers the proportion correct and the chance performance rate in estimating the proportion of discriminators. For the same proportion correct in the triangle and the 3-AFC test, there will be the same estimated proportion of discriminators since they have the same chance probability level (1/3). However, the Thurstonian/SDT model tells us that the triangle test is harder. For the same proportion correct in an experiment, there must be much better discriminability of the items to achieve that level in the triangle test. In other words, it was necessary for the products to be more dissimilar in the triangle test—since the triangle test is harder, it took a bigger difference to get to the observed proportion, as opposed to the 3-AFC. Obviously, being a "discriminator" in a triangle test requires a larger perceptual difference than being a discriminatory in 3-AFC. So the notion of discriminators is specific to the method employed. However, in spite of this logical limitation, the correction-for-guessing approach has some value in helping to make decisions about sample size, beta-risk, and power estimation. As long as one is always using the same test method, the problem of different $d'$-values need not come into play.

The use of $d'$ as a criterion has an important limiting factor as well. The variance of a $d'$ value is given as the value of a $B$-factor divided by $N$, the number of judges or observations (ASTM, 2008b) (see Table O for values of $B$). Unfortunately, the $B$-factor passes through a minimum near a $d'$ or 2.0 and starts to increase again as $d'$ approaches zero. This makes it difficult, from any practical perspective, to find a significant difference between some $d'$ that you might choose as an acceptable upper limit and a low level of $d'$ that you may find in the test you perform. For all practical purposes, testing an obtained $d'$ against a $d'$ limit less than 1.5 is not very efficient and demonstrating that a $d'$ is significantly lower than 1.0 is very difficult given the size of most discrimination testing panels ($N = 50$–$100$). For this reason, conclusions about similarity using $d'$ need to be based on simple rules-of-thumb, for example, by comparing the level of $d'$ to those that have previously been found to be acceptable (see ASTM (2008b) for further discussion).

In conclusion, we offer the following guidelines for those seeking evidence of sensory equivalence or similarity: First, apply the common sense principles discussed at the beginning of this chapter. Make sure you have a sensitive test instrument that is capable of detecting differences. If possible, include a control test to show that the method works or be prepared to illustrate the track record of the panel with previous results. Second, do power and sample size calculations to be sure you have an adequate panel size and adequate appreciation of what the test is likely to detect or miss. Third, get management to specify how much

of a difference is acceptable. A company with a long history of difference or equivalence testing may have a benchmark $d'$, a proportion of discriminators or some other benchmark or degree of difference that is acceptable. Fourth, adopt one of the statistical approaches such as a similarity test, interval testing (see Ennis and Ennis, 2010, for a new approach), or TOST to prove that you are below (or within) some acceptable limit of variation. Finally, be aware of the power of your test to detect a given degree of difference. The best measures of degree of difference from choice tests are given by the Thurstonian delta or $d'$ values which are independent of the particular test method.

## Appendix: Non-Central *t*-Test for Equivalence of Scaled Data

Bi (2007) described a similarity test for two means, as might come from some scaled data such as acceptability ratings, descriptive panel data, or quality control panel data. The critical test statistic is $T_{\mathrm{AH}}$ after the original authors of the test, Anderson and Hauck. If we have two means, $M_1$ and $M_2$, from two groups of panelists with $N$ panelists per group and a variance estimate, S, the test proceeds as follows:

$$T_{\mathrm{AH}} = \frac{M_1 - M_2}{s\sqrt{2/N}} \tag{5.13}$$

The variance estimate, $S$, can be based on the two samples, where

$$S^2 = \frac{S_1^2 + S_2^2}{N} \tag{5.14}$$

and we must also estimate a non-centrality parameter, $\delta$,

$$\delta = \frac{\Delta_{\mathrm{o}}}{s\sqrt{2/N}} \tag{5.15}$$

where $\Delta_{\mathrm{o}}$ is the allowable difference interval.

The calculated $p$-value is then

$$p = t_\nu(|T_{\mathrm{AH}}| - \delta) - t_\nu(-|T_{\mathrm{AH}}| - \delta) \tag{5.16}$$

and $t_\nu$ is the $p$-value from the common central $t$-distribution value for $\nu = 2(N–1)$ degrees of freedom. If $p$ is less than our cutoff, usually 0.05, then we can conclude that our difference is within the acceptable interval and we have equivalence.

For paired data, the situation is even simpler, but in order to calculate your critical value, you need a calculator for critical points of the non-central $F$-distribution, as found in various statistical packages.

To apply this, perform a simple dependent samples (paired data) $t$-test. Determine the maximum allowable difference in terms of the scale difference and normalize this by stating it in standard deviation units. The obtained value of $t$ is then compared to the critical value as follows:

$$C = \sqrt{F} \tag{5.17}$$

where the $F$ value corresponds to a value for the non-central $F$-distribution for 1, $N$–1 degrees of freedom, and a non-centrality parameter, given by $N(\varepsilon)$, and $(\varepsilon)$ is the size of the critical difference in standard deviation units. If you do not have easy access to a calculator for the critical values of a non-central $F$, a very useful table is given in Gacula et al. (2009) where the value of $T$ may be directly compared to the critical value based on an alpha level of 0.05 and various levels of $(\varepsilon)$ (Appendix Table A.30, pp. 812–813 in Gacula et al., 2009). The absolute value of the obtained $t$-value must be less than the critical $C$ value to fall in the range of significant similarity or equivalence.

Worked examples can be found in Bi (2005) and Gacula et al. (2009).

## References

ASTM. 2008a. Standard guide for sensory claim substantiation. Designation E-1958-07. Annual Book of Standards, Vol. 15.08. ASTM International, West Conshohocken, PA, pp. 186–212.

ASTM. 2008b. Standard practice for estimating Thurstonian discriminal differences. Designation E-2262-03. Annual Book of Standards, Vol. 15.08. ASTM International, West Conshohocken, PA, pp. 253–299.

Amerine, M. A., Pangborn, R. M. and Roessler, E. B. 1965. Principles of Sensory Evaluation of Food, Academic Press, New York, pp. 437–440.

Antinone, M. A., Lawless, H. T., Ledford, R. A. and Johnston, M. 1994. The importance of diacetyl as a flavor component in full fat cottage cheese. Journal of Food Science, 59, 38–42.

Baird, J. C. and Noma, E. 1978. Fundamentals of Scaling and Psychophysics. Wiley, New York.

Bi, J. 2005. Similarity testing in sensory and consumer research. Food Quality and Preference, 16, 139–149.

Bi, J. 2006a. Sensory Discrimination Tests and Measurements. Blackwell, Ames, IA.

Bi, J. 2006b. Statistical analyses for R-index. Journal of Sensory Studies, 21, 584–600.

Bi, J. 2007. Similarity testing using paired comparison method. Food Quality and Preference, 18, 500–507.

Byer, A. J. and Abrams, D. 1953. A comparison of the triangle and two-sample taste test methods. Food Technology, 7, 183–187.

Delwiche, J. and O'Mahony, M. 1996. Flavour discrimination: An extension of the Thurstonian "paradoxes" to the tetrad method. Food Quality and Preference, 7, 1–5.

Ennis, D. M. 1990. Relative power of difference testing methods in sensory evaluation. Food Technology, 44(4), 114, 116–117.

Ennis, D. M. 1993. The power of sensory discrimination methods. Journal of Sensory Studies, 8, 353–370.

Ennis, D. M. 2008. Tables for parity testing. Journal of Sensory Studies, 32, 80–91.

Ennis, D. M. and Ennis J. M. 2010. Equivalence hypothesis testing. Food Quality and Preference, 21, 253–256.

Ennis, D.M. and Mullen, K. 1986. Theoretical aspects of sensory discrimination. Chemical Senses, 11, 513–522.

Ennis, D. M. and O'Mahony, M. 1995. Probabilistic models for sequential taste effects in triadic choice. Journal of Experimental Psychology: Human Perception and Performance, 21, 1–10.

Ferdinandus, A., Oosterom-Kleijngeld, I. and Runneboom, A. J. M. 1970. Taste testing. MBAA Technical Quarterly, 7(4), 210–227.

Finney, D. J. 1971. Probit Analysis, Third Edition. Cambridge University, New York.

Frijters, J. E. R. 1979. The paradox of the discriminatory nondiscriminators resolved. Chemical Senses, 4, 355–358.

Frijters, J. E. R., Kooistra, A. and Vereijken, P. F. G. 1980. Tables of d′ for the triangular method and the 3-AFC signal detection procedure. Perception and Psychophysics, 27(2), 176–178.

Gacula, M. C., Singh, J., Altan, S. and Bi, J. 2009. Statistical Methods in Food and Consumer Research. Academic and Elsevier, Burlington, MA.

Green, D.M. and Swets, J. A. 1966. Signal Detection Theory and Psychophysics. Wiley, New York.

Lawless, H. T. 2010. A simple alternative analysis for threshold data determined by ascending forced-choice method of limits. Journal of Sensory Studies, 25, 332–346.

Lawless, H. T. and Schlegel, M. P. 1984. Direct and indirect scaling of taste—odor mixtures. Journal of Food Science, 49, 44–46.

Lawless, H. T. and Stevens, D. A. 1983. Cross-adaptation of sucrose and intensive sweeteners. Chemical Senses, 7, 309–315.

Macmillan, N. A. and Creelman, C. D. 1991. Detection Theory: A User's Guide. University Press, Cambridge.

MacRae, A. W. 1995. Confidence intervals for the triangle test can give reassurance that products are similar. Food Quality and Preference, 6, 61–67.

Meilgaard, M., Civille, G. V. and Carr, B. T. 2006. Sensory Evaluation Techniques, Fourth Edition. CRC, Boca Raton.

Morrison, D. G. 1978. A probability model for forced binary choices. American Statistician, 32, 23–25.

O'Mahony, M. A. 1979. Short-cut signal detection measures for sensory analysis. Journal of Food Science, 44(1), 302–303.

O'Mahony, M. and Odbert, N. 1985. A comparison of sensory difference testing procedures: Sequential sensitivity analysis and aspects of taste adaptation. Journal of Food Science, 50, 1055.

O'Mahony, M., Masuoka, S. and Ishii, R. 1994. A theoretical note on difference tests: Models, paradoxes and cognitive strategies. Journal of Sensory Studies, 9, 247–272.

Schlich, P. 1993. Risk tables for discrimination tests. Food Quality and Preference, 4, 141–151.

Stillman, J. A. and Irwin, R. J. 1995. Advantages of the same-different method over the triangular method for the measurement of taste discrimination. Journal of Sensory Studies, 10, 261–272.

Thurstone, L. L. 1927. A law of comparative judgment. Psychological Review, 34, 273–286.

Ura, S. 1960. Pair, triangle and duo-trio test. Reports of Statistical Application Research. Japanese Union of Scientists and Engineers, 7, 107–119.

USFDA. 2001. Guidance for Industry. Statistical Approaches to Bioequivalence. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). http://www.fda.gov/cder/guidance/index.htm

Viswanathan, S., Mathur, G. P., Gnyp, A. W. and St. Peirre, C. C. 1983. Application of probability models to threshold determination. Atmospheric Environment, 17, 139–143.

Welleck, S. 2003. Testing Statistical Hypotheses of Equivalence. CRC (Chapman and Hall), Boca Raton, FL.